Socialist Federal Republic of Yugoslavia

FEDERAL PATENT OFFICE

[Coat of Arms]

The RO* "Center for Genetic Engineering", Scientific Institute for Molecular Biology, Vojvode Stepe [street] 283. 11000 Belgrade.

has filed with the Federal Patent Office a patent application requesting patent protection in the Socialist Federal Republic of Yugoslavia for an invention related to:

"PROCEDURE FOR SEQUENCING GENOMES BY HYBRIDIZATION WITH OLIGONUCLEOTIDE PROBES"

This application with enclosures thereto was received on April 1, 1987 and was registered under Number P-570/87.

The Federal Patent Office hereby certifies that the enclosed copy of said application and of the enclosures thereto conform to the original.

Granted by the Federal Patent Office, after receiving payment of the correct fee, on February 16, 1988 under No. 1932/88 in Belgrade.

[Wax seal]

[Rubber stamp:]
Federal Patent Office.
Belgrade

By authorization of
the Director
[Signed:]
Rada Vukovic

*) Translator's Note: RO = Radnicka Organizacija = Workers' Organization

RO CENTER FOR GENETIC ENGINEERING

GENOME SEQUENCING PROCEDURE BY HYBRIDIZATION WITH OLIGONUCLEOTIDE PROBES

a) Technical Field

The present invention is in the field of molecular biology. In the international patent classification, it belongs in class.....

b) Technical Problem

The size of genomes ranges from about 4×10^6 base pairs (bp) in *E. coli* to 3×10^9 bp in mammals. Determining the primary structure or the sequence of the entire genome, particularly the human genome, is our challenge at the end of the 20th century. An even greater challenge for biology is the determination of the entire genomic sequence for characteristic species of the living world. This would provide a qualitative jump in the interpretation of the functioning and evolution of organisms. It would also represent a major jump in the explanation and curing of many diseases, in food production and in biotechnology in general.

c) State of the Art

The technology of recombinant DNA has made it possible to replicate and isolate short fragments of genomic DNA (from 200 to 50,000 bp). In this manner, a sufficient amount of material was obtained for determining the sequence in which the nucleotides in the cloned fragment are arranged. The sequence is determined on polyacrylamide gels capable of separating DNA fragments of 1 to a maximum of 500 bp and differing by the length of one nucleotide. The four nucleotides are differentiated in two ways: by specific chemical degradation of the DNA strand at sites where the particular nucleotide is located, by the Maxam-Gilbert method [Maxam, A.M., and Gilbert, W., Proc. Natl. Acad. Sci. 74, 560 (1977)], and by using enzymatic DNA synthesis on the cloned matrix which involves the addition of a dideoxynucleotide capable of stopping the synthesis at all sites at which this nucleotide is located in the cloned fragment, by the method of Sanger [Sanger, F., et al., Proc. Natl. Acad. Sci. 74, 5463 (1977)]. Both methods require a considerable amount of manual work so that the rate of sequencing in good laboratories throughout the world is about 100 bp per day per person. By use of electronics (computers and robots), sequencing can be accelerated by a few orders of magnitude. The idea of sequencing the entire human

genome has been discussed at many scientific meetings in the United States [Science 232, (Research News), 1598-1599 (1986)]. The general conclusion is that sequencing can be accomplished only in well organized centers (sequencing factories), that the cost would be about 3 billion dollars and that the task would take at least 10 years. Japanese experts are currently ahead of all others in organizing components of such a center. Their sequencing center has a capacity of about one million bp per day, the cost being 0.17 dollar per genomic bp [Nature 325, (Commentary), 771-772 (1987)]. Because the random selection of cloned fragments containing about 500 bp requires sequencing three genome lenghts, the sequencing of 10 billion bp in such a center would take 30 years, namely to sequence the human genome alone in a few years, at least 10 such centers would be needed.

d) Description

Our sequencing procedure has an entirely different logic and is applicable only to the determination of sequences of the entire genome(s): it is uneconomical for the determination of specific short fragments. The procedure is based on strictly specific hybridization of oligonucleotide probes (ONPs) that are 10 to 40 nucleotides long. Because hybridization conditions can be determined when ONPs hybridize only to sequences with complete homology, the sequence can be read by such hybridization. By hybridizing the entire genomic DNA replicated in fragments of appropriate length with a sufficient number of ONPs and by computerized arrangement of the detected sequences, the entire genome can be sequenced at the same time. We believe that this procedure is several times faster and less expensive than the procedure now being developed and that for this reason it could be applicable to the sequencing of genomes of all characteristic species.

For this procedure, it is necessary to optimize the length, sequence and number of the ONPs, the length of the genomic DNA fragments that represent a hybridization point, and the method of separate replication of each such fragment.

The number of possible arrangements of the four nucleotides as a function of length is equal to 4^{length} and for some lengths is shown in the following table

Length (bp)	9	10	11	12	13
Number	262144	1048576	4194304	16777216	67108864

On the basis of the foregoing, to detect every possible sequence, namely to accomplish the displacement of only one bp in the ONP arrangement, it is necessary to use about 260.000 9-mers to 67×10^6 13-mers. Because specific hybridization is likely to be achieved only with an ONP with 10 nucleotides or more [Wallace, R.B., Nucleic Acids Research 6, 3543-3557 (1979)], the number of required ONPs will be smallest if a 10-mer or 11-mer

is used. Because of the existence of two complementary DNA strands, one ONP detects two different sequences when read in a single sense. For example, the 5'CAGA3' also detects the sequence 5'TCTG3', namely it detects

5'CAGA3'	and	5'TCTG3'
3'GTCT5'		3'AGAC5'

For this reason, only one half as many ONPs, namely a maximum of about 2 million 11-nucleotide ONPs, are needed. Palindromic ONPs are an exception. Among all 11-mers, 4^6, namely 4094 or only one thousandth, are palindromic. Conversely, this means that the frequency of nonpalindromic probes in a genome is twice as high.

For unequivocal sequence determination, it is not necessary to utilize all ONPs of a given length. The use of a smaller genomic fragment as a hybridization point makes it possible to use fewer probes. In this case, the probe overlap will be less but still sufficient so that in the short genomic DNA fragment sequences of overlap lengths will not be repeated many times.

From the average distance (S) between sites complementary to one ONP, which depends on the ONP length and the ratio of its dinucleotide composition to the dinucleotide composition of the genomic DNA being sequenced, it is possible to determine the frequency of the given sequence along a certain length of genomic DNA. We used equations derived on the basis of the theory of probability [Drmanac, R., et al., Nucleic Acids Research 14, 4691-4692 (1986) and our Patent Application No. 5742 of March 24, 1987]. Table 1 shows the average distance between sequences of certain homologous ONPs in mammalian genomes.

Table 1

Average Distance (S) Between Sequences of Homologous ONPs in Mammalian Genomes

	ONP Length (bp)				
C + G	7	8	9	10	11
0	2300	7600	25400	85000	282000
1	3450	11400	38100	127500	423500
2	5170	17100	57150	191250	634000
3	7750	25600	85700	285000	951000
4	11600	38500	128600	330000	1427000
5	17500	57700	190000	495000	2140000
6	26200	86600	285000	742000	3210000
7	39300	129900	330000	1113000	4816000
8		195000	495000	1670000	7724000

From the average distance (S) and using the following equation, we calculated the percentage of genomic fragments of length D within which the sequence recognized by the given ONP is repeated at least once:

$$P(D) = [1 - (1 - 1/S)^D] \times 100 \qquad (1)$$

The results are presented in Table 2.

Table 2

Percentage of 5000 to 20,000 bp-long Genomic DNA Fragments Containing Sequences of Complementary ONP with S Equal to 25,000 to 200,000 bp

	S			
D	25000	50000	100000	200000
5000	18	9.5	5	2.5
10000	32	18	9.5	5
20000	55	33	18	9.5

By using the binomial distribution, we determined the probability that the sequence that is complementary to the given ONP will be repeated a certain number of times in a DNA fragment of defined length. This probability depends on the average distance (S) separating the given sequence in the genome. This probability can be calculated by the following equation:

$$P(N) = C(D,N) \times (1/S)^N \times (1 - 1/S)^{D-N} \qquad (2)$$

wherein D is the length of the DNA fragment in bp and N is the number of repetitions of the given sequence within the length D whose probability P(N) is being sought. C(D,N) is the number of combinations without repetition of class N of D elements. Because within length D there are approximately D sequences which on average have the same S, by multiplying the above probability by D we obtain the number of different sequences of a defined S that are repeated N times within length D. The calculated numbers of different sequences within a given range of S, D and N are presented in Table 3.

Table 3

Number of Oligonucleotide Sequences with S in the Range of 25,000 to 200,000 bp and Repeated 2 and 3 Times in 5000 bp and 10,000 bp Fragments

		S			
D	N	25000	50000	100000	200000
5000	2	82	22	4.5	1.6
	3	5	0.75	0.07	0.013
10000	2	558	163	45	11
	3	74	10	1.5	0.18

From Table 3 it is possible to estimate the minimum required number of ONPs of a defined length and C+G composition necessary for successful reading of sequences in the 5 kbp or 10 kbp long fragment. The frequency of two kinds of sequences is essential for such reading, namely the frequency of sequences of complementary ONPs and that of the sequences of ONPs with average maximum ONP overlap. This will be explained on the example of a nucleotide with 11 ONPs.

For about 2×10^6 ONPs of 11 bp length, every 11-mer is detected in any DNA sequence. In this case, overlap is always at a maximum and amounts to 10 bp. It can be seen from Table 1 that the average distance between the most frequent 11-mers is 282,000. If all 4989 11-mers in the 5000 bp fragment were equally frequent, only one of them would be likely to be repeated twice (Table 2). Because there are probably n sequences of 5000 bp length

with 90% or more of A + T, this means that a more significant repetition of 11-mers in the 5000 bp fragment and probably also in the 10,000 bp fragment can occur only for nonrandom reasons. With regard to the overlap sequence (10-mers), the highest frequency occurs for an average of 85,000 (Table 2), and 2-fold repetition within 5000 bp would occur for a maximum of five such sequences, and probably for an average of about two sequences.

The ease and accuracy, namely the nonambiguity, of the reading of sequences depends on the number of repetitions of overlapping sequences. If we imagine reading as a two-dimensional progression of one or more starts (randomly selected ONPs from among all ONPs capable of hybridization to the given genomic DNA fragment), then for each starting 11-mer we look for the left and the right base pair by searching among the hybridized ONPs for the 10-mer that is to the left and the right of the starting 11-mer. When after a certain number of reading steps the 10-mer is found which because of being repeated in the given sequence is present in more than one 11-mer, the reading in this sense must be interrupted here, because we do not know which of the detected base pairs are in the continuation of the sequence and which are at some other location. By reading in the other sense, this interruption will be overcome. Considerable repetition of overlapping sequences, however, will make reading more difficult, and it may even become impossible to overcome the interruption.

On the basis of the calculated repeatability, it is possible to estimate the lowest number of 11-nucleotide ONPs required to prevent the interruption of sequence reading or the ambiguous linking of the read fragments. By reducing the number of ONPs, the overlapping sequence is shortened and its repeatability is thus increased. By synthesizing a larger number of more frequent 11-mers (containing more A and T) and a lesser number of those with more C and G, it is possible to achieve the same optimal repeatability of overlapping sequences although of different lengths. Assuming that the maximum repeatability of overlapping sequences resulting in successful reading is about 20 sequences repeated twice, for the sequencing of 5000 bp fragments, the average distance between overlapping sequences must not be less than 50,000 bp. This means that the following needs to be synthesized: all ONPs with one or without any C or G (this gives an overlap length of 10 bp); every other 11-mer with C + G from 2 to 4 (this gives an overlap length of 9 bp); every third 11-mer with C + G from 5 to 7 (this gives an overlap length of 8 bp), and every fourth 11-mer with C + G greater than 7 (this gives an overlap length of 7 bp). The total number of ONPs thus selected would be about 8×10^5. In our opinion, computer simulation would show that even one half of this number of 11-nucleotide ONPs would be sufficient. The sequencing of 10,000-bp fragments would require about 10^6 11-mers. If 12-mers were used, this number would be at least three times higher.

For easier reading, synthetic ONPs can be arranged by starting from one or several ONPs and proceeding over the overlapped parts. The ONPs thus arranged would be marked by letters in alphabetical order and according to increasing numbers. Such marking would make it possible to arrange the ONPs that hybridize to a given genomic DNA fragment into one or several arrays which would then be converted to the DNA sequence only by deciphering.

The replication of genomic DNA in fragments of defined length can be accomplished in two ways: 1) by cloning, and 2) by amplification.

It can be seen from the foregoing analysis that the maximum length that can be read with a reasonable number of ONPs is about 10,000 bp and that 4000-5000 bp is a better length. Plasmid vectors are most advantageous for cloning these lengths. To create a complete genomic library, these vectors, because of their lower transformation efficacy compared to phage vectors, require 20 to 100 μg of genomic DNA, which is not a major requirement for the one-time creation of the library. For a better representation of genomic DNA, it would be necessary to generate 5000 bp long fragments by partial digestion with two to three common enzymes (Sau 3A, Ddel, Alu I). To reduce the effect of any "toxic" and repetitive sequences (in this respect, plasmid vectors have an advantage over phage or cosmid vectors), it is necessary to form a library in two vectors. In our opinion, plasmids of series pUC and pAT are most advantageous for this purpose because they multiply well and are relatively small.

The sequencing of cloned fragments by hybridization can be accomplished in two ways: by colony hybridization and by dot blot hybridization of isolated plasmid DNA. In both cases, 2000 to 3000 different ONPs represented in the vector sequence cannot be utilized, i.e., they will not even be synthesized.

Colony hybridization is probably faster and less expensive than dot blot hybridization, but it requires specific conditions to eliminate the effect of hybridization with bacterial DNA. To reduce general background noise, the labeling of probes should confer high sensitivity in hybridization, because in this manner very small colonies could be used. ONP labeling should in any case be by biotinylization because of easy and lasting labeling in the last synthesis step. The sensitivity achieved in this case [Al-Hakin, A.H., and Hull, R., Nucleic Acid Research 14, 9965-9976 (1986)] makes it possible to utilize at least 10 times fewer colonies than are required by the standard method.

To avoid false positive hybridizations caused by homology of the ONP with the bacterial sequence and to utilize short probes such as the 11-mers, which on average are repeated twice in the bacterial chromosome, it is necessary to use vectors giving a maximum number of copies per cell. It is known that by additional amplification on chloramphenicol, pBR 322 can produce 300 to 400 copies per bacterial cell [Lin Chao, S., and Bremer, L., Mol. Gen. Genet. 203, 150-153 (1986)]. The replication efficacy of the plasmids pAT and pUC is at least twice as high [Twigg, A.J. et al., Nature 283, 216-218 (1980)], so that we can assume that under optimum conditions even 500 plasmid copies can be produced per cell. Because of the load represented by the sequence introduced, the chimeric plasmids will certainly not multiply as well, particularly in the presence of more toxic sequences. For this reason it is necessary to work with about 200 copies of chimeric plasmid per cell. This means that, on average, with each

11-mer the signal would be 100 times stronger if the complementary sequence were located on the plasmid. This represents a sufficient difference so that with a small amount of DNA, namely by use of small hybridization colonies, hybridization with the bacterial DNA would not register.

By using the binomial distribution, we determined how many ONPs will be repeated in the bacterial chromosome more than 10 times as a result of random distribution. Such ONPs would give unreliable information or, if they gave approximately the same signal strength with all colonies, they could not be used at all.

Table 3 shows the results obtained by use of Eq. 2, wherein D is the length of the bacterial chromosome, i.e., 4×10^6 bp, and S is the number of different ONPs. This calculation assumes that all nucleotides and dinucleotides are uniformly represented in the DNA of *E. coli*, which is almost entirely the case.

Table 4

Probability of a Given 11-mer Frequency

in the *E. coli* Genome

No. of repetitions (n)	0	2	4	6	8	10	14
Percent 11-mers	13.5	27	9	1.2	0.086	0.004	7x10-6
Total no. of 11-mers	-	-	-	-	1720	80	0.14

It can be seen from Table 4 that it cannot be expected that any 11-mer will be repeated more than 13 times, and that 300 is the total number of those that are repeated more than 10 times. This means that the vast majority of 11-mers will have a more than 20 times stronger signal originating from the cloned DNA than from the bacterial DNA. The naturally determined number of 11-mers will for functional reasons be large in bacterial DNA, but because, as a result of recombination, bacteria do not tolerate significant repetition, we can expect that the number of such 11-mers will be small. They simply would not be utilized for hybridization.

The problem of hybridization with bacterial DNA can also be solved by selective prehybridization using "cold" bacterial DNA. By preparing this DNA in fragments larger than 100 bp and smaller than about 10,000 bp under stringent hybridization conditions in which only fragments with homology greater than 50 bp undergo hybridization, bacterial DNA would be preferentially "covered". This is because the probability that there are random homologous sequences of 50 bp or longer between bacterial and eukariotic DNA is negligible.

Selective prehybridization also makes it possible to use several probes simultaneously in the colony hybridization sequencing procedure. In this manner, the required number of independent hybridizations can be

reduced. On the other hand, to determine which ONP or ONPs enable the combination to undergo positive hybridization, each probe must be present in several combinations, and this increases the required amount of each ONP. However, because for successful and fast hybridization it is necessary to achieve a certain concentration of probes in the hybridization liquid, and because probe consumption is very low so that the concentration is only very slightly reduced after hybridization, a larger number of filters can be hybridized in a few portions of smaller volume of the same hybridization liquid, which requires a smaller amount of probe or probes.

By using 30 ONPs per hybridization and by repeating one ONP in three combinations so that none of the other 90 probes is present in two of the three combinations, the number of hybridizations is reduced tenfold at the cost of a three times larger amount of each ONP required. Based on the probability that the combination of a defined number of ONPs hybridizes to the fragment of genomic DNA of defined length, we determined the percentage of information that is lost compared to when each ONP is used separately.

The average distance between homologous sequences for 30 ONPs with 11 nucleotides is about 130,000 bp. For the sequencing of mammalian genomes, because of the more accurate reading, proportionally more ONPs with a more frequent homologous sequence, namely containing more A and T bases, would be synthesized. Hence, in this case, the average distance (S) would be about 100,000 bp. By use of the equation $P(D) = 1 - (1 - 1/S)^0$ (Eq. 1), we determined the probability that a combination of 30 ONPs will hybridize to a genomic DNA fragment of length D = 5000 bp. This probability is 0.0485. The probability that three different combinations will hybridize to the same fragment is 1.25×10^{-4}. Since 2 million colonies (fragments) are being hybridized, in about 250 colonies all three combinations that have a common ONP will hybridize to at least one of their probes. For these colonies, we will not know whether they have a sequence complementary to the common ONP. Because for mammalian genomes the number of clones that contain at least one complementary ONP sequence that is common to the three combinations is 300 to 30,000, the number of colonies that will also simultaneously hybridize with the common probe will in the worst case be less than four. For one million different ONPs this means a maximum of 4 million lost pieces of information assuming that the common ONP does not hybridize wherever there is ambiguity as to whether it hybridizes or not. This represents a loss of only one millionth part of the information that would be obtained by hybridization with each ONP separately.

Information is more likely to be lost by rejection of positive hybridization with the ONP that is common to the three combinations that hybridize to the given genomic fragment as a result of erroneous determination that each combination contains at least one ONP that hybridizes with the given fragment. This error arises in the determination of positive hybridization for each ONP from the three combinations involved when one considers whether the other two combinations that contain them also hybridize. If the other two combinations hybridize, then a high probability exists that the positive hybridization is due to the common ONP. On the other hand, if the

combinations are large, the probability is higher that two different probes undergo hybridization, each in one combination. This would mean that the common ONP probably does not hybridize and, hence, we would not have to reject the initially considered ONP as the one that does not hybridize to the given fragment. This probability (Pgi) can be calculated approximately by use of the equation $Pgi = \{[(D)]^2 \times K\}^3$ wherein K is the number of ONPs in the combination and P(D) is the probability that at least one of K ONPs hybridizes to one fragment of genomic DNA having length D (Eq. 1). The formula is valid for $[P(D)^2] \times K < 1$. When fragments of length D = 5000 bp are sequenced, then with combinations having K = 30 ONP, 0.1% of the information is lost; with K = 40, ONP 0.5% is lost; with K = 45 ONP, 1.32% is lost; with K = 50 ONP, 3.3% is lost and with K = 60 ONP, 16% of the information is lost. It can be concluded that a 10-15-fold reduction in the required number of hybridizations can be achieved with a small loss of information. This time the number of required filters, namely the number of replications to be made of 2 million clones, would also be smaller.

The total number of hybridization points can be reduced by using a few hybridization steps with large combinations. Thus, at the expense of 2-3000 additional hybridizations and 2-3 rearrangements of hybridization points, each point could be searched with 3-4 times fewer hybridizations, namely the transfers to the filter could thus be reduced this many times.

By hybridization of the isolated plasmid DNA, the hybridization procedure would be facilitated, but it would be necessary to isolate a sufficient quantity of plasmid DNA from many clones. The number of clones with 5000-bp fragments for threefold covering of mammalian genomes is 2×10^6. The required amount of DNA from each clone (Mp) is given by the product

$$Mp = (Dp/D_{ONP}) \times Bh \times (1/Br) \times Md$$

where Dp is the size of the chimeric plasmid in bp, D_{ONP} is the ONP length, Bh is the number of required hybridizations, Br is the number of rehybridizations of the same filter and Md is the amount of DNA that can be detected by the hybridization procedure. By taking the most probable values, namely Dp = 8000, D_{ONP} = 11, Bh = 2×10^4, Br = 10 and Md = 0.1 pg, we find that it is necessary to isolate about 0.2 μg of DNA for each chimeric plasmid. Successful rehybridization of filters that have been hybridized with a biotinylized probe has not been developed to date. On the other hand, there are indications that with biotinylized probes it is possible to detect as little as 0.001 pg. Hence, from each of the 2×10^6 clones it is necessary to isolate about 0.1 to 1 μg of plasmid DNA.

The amplification of the entire genomic DNA can be accomplished in about one million portions of a size up to 10,000 bp whereby the genome would be covered more than three times. This is accomplished by means of an

appropriately chosen mixture of oligonucleotides as primers (our Patent Application No. 5742 of March 24, 1987). With about 50,000 different oligonucleotides having the complementary sequence repeated 800 times in the nonrepetitive part of the mammalian genome (for example, a 12-mer with C + G from 1 to 5), it is possible to carry out one million amplification reactions with combinations containing 50 primers so that each primer enters only once into the same combination with every other primer. With such primer combinations, there will be an average of 60 sites in the genome where two primers will be oriented so that their 3' ends will face each other and will be separated by a distance of less than 300 bp. The fragments between these primers will be amplified. Because their average length is 150 bp, the total length of the amplified genome is about 9000 bp. One million of such amplification reactions replaces the plasmid and phage library of the mammalian genome. In the amplification it is not possible to utilize primers that enter into highly repetitive sequences (those that are repeated more than 2-3000 times); hence only the amplification or sequencing of the nonrepetitive part of the genome takes place. In addition, with 50,000 primers with a frequency of 800 in the nonrepetitive part of the genome, about 10% of this part of the genome would not enter into the amplification units. With 100,000 primers, only 0.1% of the nonrepetitive part of the genome would remain unamplified. With 100,000 primers, it is necessary to carry out 4 million amplification reactions.

By dot blot hybridization of amplifying reactions with oligonucleotides that served as primers and with newly synthesized ONPs up to the required number of about one million, only the sequences of the amplified fragments would be read, because with a $2x10^4$-fold amplification each ONP having a complementary sequence in the amplified fragment would have a 3-1000 times larger number of targets than if it hybridized only to the homologous sequences in the unamplified part of the genome. Only a three times stronger signal is expected for 11-nucleotide ONPs that do not contain C or G, and a 1000 times stonger signal for 12-mers without A or T. It can be seen from this analysis that by sequencing regions rich in A and T it is possible to utilize ONPs longer than 11 bp (the 12-mer would give a signal 10 times stronger than the background noise). In this case, it is impossible to utilize ONP combinations for hybridization, because the signal would be equal to the background noise, and no possibility exists for selective prehybridization.

The advantage of amplification over cloning is that no living material is used. This procedure is much more expensive, however, because each primer is consumed in 10 times larger quantity than if it were used only as a probe. Moreover, about 10^7 to 10^8 enzyme units of the the Klenow fragment of polymerase I are required.

Because each genomic DNA fragment hybridizes with all probes, it is necessary, if there is no rehybridization and if probe combinations are not used for hybridization, to apply each colony or each isolated DNA or amplification reaction to about one million filters for about one million probes. This would be done by simultaneous automatic application of a large number of samples (about 100). With DNA, this is much easier than by making

colony replicas. Most likely, the colonies would be automatically seeded into about one million replicas of each of the approximately 2 million clones by taking a minimum amount of bacteria from clones grown on microtitration plates. To avoid removing colonies from Petri dishes, the transformation mixture can be diluted by seeding the specified volume into a hole of the microplate so that one or no transformed cell is seeded. To eliminate empty holes, a transplantation would then be performed from holes with viable growth to a new plate. The most difficult condition is the need to achieve approximately the same growth of all colonies on all filters.

If there is no rehybridization and if probe combinations are not used, the total number of hybridization points equals the product of the number of genomic DNA fragments (colonies, clones, amplification reactions) by the number of ONPs. For mammalian genomes this amounts to about 10^{12} points. If each point requires about 3 mm^2, about 3×10^6 m^2 of filters will be needed. With 10 rehybridizations and a 20-fold reduction in the number of hybridizations per fragment, about 15,000 m^2 of filters is required.

Hybridization with all ONPs of the same length would be carried out at the same temperature under conditions that eliminate the effect of the C ÷ G composition [Wood, W., et al., Proc. Natl. Acad. Sci. 82, 1585 (1985)]. For 11-mers, the hybridization and the washing would be carried out at 20 °C. For biotinylized probes, which require about 2 ng of probe per cm^2 of filter, an amount of one to three optical units of each ONP (50 μg) would be sufficient for the sequencing of a mammalian genome provided the hybridization liquid is used only once. By simultaneous synthesis of 10 optical units and possibly by simultaneous hybridization, the sequencing of individual genomes could be simplified, accelerated and made less expensive.

The cost of sequencing per genome the size of a mammalian genome would not exceed 100 million dollars. This is 5 times less expensive than the costs estimated within the framework of the Japanese project. We also believe that the total time needed for the sequencing of a genome including ONP synthesis is shorter and amounts to about 1-2 years.

Because as many genomic fragments are taken for sequencing as are necessary for each fragment to overlap the neighboring one at least slightly, from the sequenced fragments one obtains by overlapping over homologous sequences at the ends of the fragments an arranged library of fragments (clones) and the sequences of each chromosome. This is not so in the sequencing of amplified fragments, because it is possible to amplify and to sequence only fragments that do not contain, or do not belong to, repetitive sequences. In this case, by arranging the sequenced fragments, one would obtain only regions between repetitive neighboring sequences.

It appears that the optimum procedure for sequencing genomes by the method of hybridization with ONPs is by colony hybridization of clones larger than 5000 bp with about 300.000 to 500.000 ONPs with a length of 10-11 nucleotides in about 50.000 separate hybridizations with combinations of about 30-50 ONPs so distributed that each ONP is repeated in three combinations wherein the other 90 ONPs are represented only in one of the three given combinations. The possibility of detection of a very small amount of DNA by use of biotinylized probes, by increasing the number of rehybridization of a filter and by reducing the number of hybridizations per fragment in elimination hybridizations with combinations of about 100-500 ONPs, however, makes it possible to reduce the required amount of DNA to less than 1 μg. This quantity of plasmids can be isolated from bacterial cultures grown in one hole of a microtitration plate. We can also visualize simple, crude isolation of plasmid DNA, which could even be easier than growing colonies in thousands of replications. The entire isolation procedure would be carried out on microtitration plates. Centrifugation in the microtitration plates would remove the medium, and alkaline lysis and denaturation of the protein with acidic sodium acetate would give the cell membrane chromosomal precipitate, which would be removed by centrifugation. The supernatant would be denatured with sodium hydroxide and transferred to the filter in the form of a sufficient number of dots. It would be easy to introduce the steps of alcoholic precipitation and treatment of the preparations with the RNA-se enzyme if it were necessary to reduce the background noise from the hybridization with bacterial RNA. This method of isolation of plasmid DNA makes dot blot hybridization more advantageous than colony hybridization.

PATENT CLAIMS

1. The procedure of genome sequencing by hybridization with oligonucleotide probes, characterized in that genomic DNA fragments containing 100 to 20.000 bp, obtained in sufficient amount by cloning into vectors that replicate in *E. coli* of by amplification of genomic DNA with mixtures of oligonucleotide primers, or by the procedure of colony hybridization without or with selective prehybridization of nonbiotinylized bacterial DNA, or by dot blot hybridization of isolated chimeric DNA vector inserts, or by dot blot hybridization of DNA from amplification reactions, under hybridization conditions permitting only the hybridization of sequences with complete homology, are hybridized to 100.000 to 5.000.000 biotinylized oligonucleotide probes of different sequence length ranging from 10 to 13 nucleotides, each probe being hybridized separately or in combinations of 10 to 500 probes and that the groups of oligonucleotide sequences that are located in individual genomic DNA fragments are determined by detection of the bound biotin, the arrangement of said fragments over overlapping sequences then giving the order of the nucleotides in said fragments.

2. The procedure according to Claim 1, characterized in that one deduces from the combinations of oligonucleotide probes that hybridize to one fragment the oligonucleotide probes that produce hybridization by elimination of those probes whose other combinations in which they are present do not hybridize to the given fragment or that in all combinations containing the given probe and which hybridize to the given fragment there is present at least one additional probe all combinations of which that contain it hybridizing to the given fragment.

3. The procedure according to Claims 1 and 2, characterized in that the oligonucleotide probes that hybridize to the given genomic DNA fragment arrange themselves into one or several arrays in the alphabetical order of the lettered part and according to increasing value of the numbered part of their marking which they acquired on the basis of possible overlap with the utilized probes, and that the arrays of markings are deciphered by means of a reverse algorithm to give the order of nucleotides.

4. The procedure according to Claims 1 through 3, characterized in that by detecting identical, overlapping, terminal sequences between sequenced fragments, sequenced clones or amplified fragments are arranged in an array, and the overall sequence of each chromosome of the given genome is determined.

Applicant

[Signed:]

Prof. Dr. Vladimir Glisin

Director

ABSTRACT

The existence of given oligonucleotide sequences in genomic DNA fragments is determined by colony hybridization or dot blot hybridization of genomic DNA fragments with 3000 to 15.000 bp. obtained by cloning or amplification of oligonucleotide probes containing 300.000 to 3 million biotinylized probes of 10 to 12 nucleotides. individually or in combinations of 10 to 500 ONPs. under conditions permitting hybridization only with completely homologous sequences. By arranging the detected sequences over identical regions. the sequence of each individual DNA fragment is determined. By analyzing the number of fragments that cover the genome three times. each fragment can be made to overlap on both sides with at least one fragment. By detecting the fragments with identical terminal sequences. chromosome libraries and chromosomal sequences are obtained. This procedure is more than 5 times less expensive and faster than standard automated sequencing procedures.

INDUSTRIAL USE OF THE DEVELOPED GENOME SEQUENCING PROCEDURE

Based on the described sequencing procedure. it is possible to build a plant for sequencing genomic DNA. In our opinion. a large scientific-technological market will exist for genomic sequences and sequenced genomic fragments of at least 50 characteristic species.

In addition to its economic justification. such a plant would enable a large number of scientists to study the functions of genomic DNA fragments of known sequence and by procedures of genetic engineering to create new. useful combinations of genetic materials instead of cloning and sequencing certain genes at lower efficacy.

Applicant
[Signed:]
Prof. Dr. Vladimir Glisin
Director